



Division of Informatics, University of Edinburgh

**Centre for Intelligent Systems and their Applications
Institute for Adaptive and Neural Computation
Institute for Communicating and Collaborative Systems
Institute for Computing Systems Architecture
Institute of Perception, Action and Behaviour
Laboratory for Foundations of Computer Science**

informatics

by

Michael Fourman

Informatics Research Report EDI-INF-RR-0139

Division of Informatics
<http://www.informatics.ed.ac.uk/>

July 2002

informatics

Michael Fourman

Informatics Research Report EDI-INF-RR-0139

DIVISION *of* INFORMATICS

Centre for Intelligent Systems and their Applications

Institute for Adaptive and Neural Computation

Institute for Communicating and Collaborative Systems

Institute for Computing Systems Architecture

Institute of Perception, Action and Behaviour

Laboratory for Foundations of Computer Science

July 2002

entry for 'informatics' to appear in International Encyclopedia of Information and Library Science (second edition) (0415259010) John Feather and Paul Sturges eds. Routledge 2002

Abstract :

This article is an extended entry in the Routledge International Encyclopedia of Information and Library Science. It gives an account of the origins and meaning of the word 'informatics', and attempts to give some hint of the scientific depth, intellectual scope, and social importance of the subject, using examples relevant to the intended audience of this encyclopedia.

Keywords : informatics

Copyright © 2002 by Routledge

The author and the University of Edinburgh retain the right to reproduce and publish this paper for non-commercial purposes. Permission is granted for this report to be reproduced by others for non-commercial purposes as long as this copyright notice and the reference to the Encyclopedia are reprinted in full in any reproduction. Applications to make other use of the material should be addressed in the first instance to Routledge Ltd of 11 New Fetter Lane, London EC4P 4EE

The authors and the University of Edinburgh retain the right to reproduce and publish this paper for non-commercial purposes.

Permission is granted for this report to be reproduced by others for non-commercial purposes as long as this copyright notice is reprinted in full in any reproduction. Applications to make other use of the material should be addressed in the first instance to Copyright Permissions, Division of Informatics, The University of Edinburgh, 80 South Bridge, Edinburgh EH1 1HN, Scotland.

informatics

Informatics is the science of information. It studies the representation, processing, and communication of information in natural and artificial systems. Since computers, individuals and organizations all process information, informatics has computational, cognitive and social aspects.

Used as a compound, in conjunction with the name of a discipline, as in medical informatics, bio-informatics, etc., it denotes the specialization of informatics to the management and processing of data, information and knowledge in the named discipline.

Terminology

The French term *informatique*, together with various translations — *informatics* (English), *informatik* (German), and *informatica* (Italian, Spanish) — was coined by Dreyfus, in March 1962 (Dreyfus 1962), referring to the application of computers to store and process information (see also Bauer 1996). The morphology, *information* + *-ics*, uses ‘the accepted form for names of sciences, as *conics*, *linguistics*, *optics*, or matters of practice, as *economics*, *politics*, *tactics*’ (Oxford English Dictionary 1989); *informatics* encompasses both science and practice. Phonologically, *informatics* combines elements from both ‘information’ and ‘automatic’, which strengthens its semantic appeal. This new term was adopted across Western Europe, and, except in English, developed a meaning roughly translated by the English ‘computer science’, or ‘computing science’. Mikhailov *et al.* advocated the Russian term ‘informatika’ (1966), and the English ‘informatics’ (1967), as names for the ‘theory of scientific information’, and argued for a broader meaning, including study of the use of information technology in various communities (e.g. scientific) and of the interaction of technology and human organizational structures.

‘Informatics is the discipline of science which investigates the structure and properties (not specific content) of scientific information, as well as the regularities of scientific information activity, its theory, history, methodology and organization.’

(Mikhailov *et al.* 1967)

Usage has since modified this definition in three ways. First, the restriction to *scientific* information is removed, as in *business informatics* or *legal informatics*. Second, since most information is now digitally stored (Lesk, 1997; Lyman *et al.* 2000), computation is now central to informatics: Gorn (1983) defines informatics as computer science plus information science. Third, the processing and communication of information are added as objects of investigation, since they have been recognized as fundamental to any scientific account of information.

In the English-speaking world the term informatics was first widely used in the compound, ‘medical informatics’, taken to include ‘the cognitive, information processing, and communication tasks of medical practice, education, and research, including information science and the technology to support these tasks’ (Greenes and Shortliffe 1990). Many such compounds are now in use.

A June 2002 web search (Google 2002) found ‘informatics’ and various compounds, occurring more or less frequently: the numbers of documents returned for each term are given in parentheses: *informatics* (1,100,000), *bioinformatics* (691,000), *medical informatics* (151,000), *health informatics* (52,800), *museum informatics* (19,500), *nursing informatics* (15,600), *geoinformatics* (11,100), *neuroinformatics* (9,180), *social informatics* (6,840), *business informatics* (6,610), *dental informatics* (2,850), *molecular informatics* (2,630), *environmental informatics* (2,580), *legal informatics* (1,640), *chemical informatics* (1,230), *mobile informatics* (492), *protein informatics* (408), and *library informatics* (303). Some informatics specializations are named in other ways: for example, what might be called *science informatics* (1,710) is more usually called *e-science* (17,700); bioinformatics is often called *computational biology* (347,000).

Each of these areas studies representations and uses of information, which may be peculiar to each field of application, but draw on common social, logical and computational foundations. They all involve the use of computing and information technologies to store, process and communicate information. They all also address the interaction of technology with the production and use of information by individuals and organizations; they develop software, systems and services that aim to help people interact with information, efficiently and effectively.

The scope of Informatics

What these areas have in common is informatics: the focus on information and how it is represented in, processed by, and communicated between a variety of systems. Representations include paper, analogue, and digital records of text, sounds and images, as well as, for instance, the information represented in a gene, and the memories of an individual or an organization. Processing includes human reasoning, digital computation, and organizational processes. Communication includes human communication and the human-computer interface - with speech and gesture, with text and diagram, as well as computer communications and networking, which may use radio, optical or electrical signals.

Informatics studies the interaction of information with individuals and organizations, as well as the fundamentals of computation and computability, and the hardware and software technologies used to store, process and communicate digitised information. It includes the study of communication as a process that links people together, to affect the behaviour of individuals and organizations.

Informatics as a Science

Science progresses by defining, developing, criticizing and refining new concepts, in order to account for observed phenomena. Informatics is developing its own fundamental concepts of communication, knowledge, data, secrecy, interaction and information, relating them to such phenomena as computation, thought, and language, and applying them to develop tools for the management of information resources.

Informatics has many aspects. It encompasses, and builds on, a number of existing academic disciplines: primarily Artificial Intelligence, Cognitive Science and Computer Science. Each takes part of Informatics as its natural domain: in broad terms, Cognitive Science concerns the study of natural information processing systems; Computer Science concerns the analysis of computation, and the design of computing systems; Artificial Intelligence plays a connecting role, producing systems designed to emulate those found in nature. Informatics also informs, and is informed by, other disciplines, such as Mathematics, Electronics, Biology, Linguistics, Psychology, and Sociology. Thus Informatics provides a link between disciplines with their own methodologies and perspectives, bringing together a common scientific paradigm, common engineering methods and a pervasive stimulus from both technological development and practical application.

Informatics builds on a long tradition of work in logic, which provides an analysis of meaning, proof and truth. It draws on probability and statistics to relate data and information, and on the more recent tradition of computer science for abstract models of computation, and fundamental notions of computability (What can be computed?) and complexity (How do the space and time requirements of a computation scale as we consider problems of different sizes?). Combining these traditions enriches them all, since they share a common interest in information.

The science of information provides a new paradigm of scientific analysis, which concentrates on the processing and communication information, rather than focussing on the electrical, optical, mechanical or chemical interactions that embody this activity. Focussing on information provides novel accounts of long-standing phenomena, even in Physics (Frieden 1998). Informatics provides accounts of the representation, processing, and communication of information, the conceptual basis for applying computing and information technologies to develop new tools for the management of information in diverse areas, and also the basis for beginning to unravel the workings of the mind.

We illustrate the scope and nature of informatics with a brief account of one area in which informatics is contributing to library and information science. With essentially all information becoming available online, libraries will focus increasingly on selection, searching, and quality assessment. Informatics' contribution to this enterprise is to provide and apply appropriate techniques for the representation, processing and communication of information. We give an account of some of these techniques, with a focus on textual information.

Representation

'The Web was designed as an information space, with the goal that it should be useful not only for human-human communication, but also that machines would be able to participate and help. One of the major obstacles to this has been the fact that most information on the Web is designed for human consumption. [...] The Semantic Web approach [...] develops languages for expressing information in a machine processable form.'

(Tim Berners-Lee 1998)

Data on the World Wide Web makes digital representations familiar to us all. Texts, images and sounds are represented by digital encodings, as patterns of bits. Standards, such as ASCII, Unicode, JPEG, TIFF, WAV and

MP3, allow for the encoding, storage, exchange, and decoding of multimedia information. These representations may be indexed and retrieved as individual files, but they are, to varying degrees, opaque to software agents searching for information.

Text files are stored as unstructured sequences of characters. Dividing a text into words, sentences and paragraphs is mostly straightforward. Finding keywords in a text file is straightforward, and some search engines incorporate shallow linguistic knowledge that extends keyword search. For example, 'stemming' determines the morphological root of a given word form, thus relating singular and plural forms of a noun, and different moods and tenses of a verb. Searching for swan, using stemming finds swans, and searching for goose should find geese (see free text searching).

We can also search for content in other media. Finding keywords in recorded speech is much harder than searching a text file: more expensive in computing resources, and less accurate. Finding images with specified content, just by examining the image automatically, is even harder (see image retrieval).

Extracting information from texts requires deeper linguistic analysis. For example, searching news reports to glean information about company takeovers - who is being taken over, by whom, what price is being paid, for what - requires a grammatical analysis of individual sentence structures. More complex information - such as finding arguments that support a particular decision - demands more global analyses of meaning. Text files form a small fraction of society's data storage (Lyman and Varian 2000). However, texts are rich in information that is not represented elsewhere, so it is important to make this information accessible. If texts are stored in ways that make grammatical and rhetorical structures transparent, then it becomes easier for automated tools to access such information.

Metadata

The simplest way of making searching easier is to attach an electronic 'catalogue card' to each document. Such data about data is called 'metadata'. For example, the Dublin Core is a metadata standard which specifies a set of required and permitted elements for such a catalogue card. The Resource Description Format (RDF) is a general standard for such metadata.

Metadata allows software agents to find, retrieve, and process data. Just as books in a library are made accessible by the catalogue, so information on the web is made accessible by metadata.

The Semantic Web is a project that aims to provide a common framework for such efforts, by having data on the web defined and linked in such a way that it can be used by machines not just for display purposes, but for automation, integration and reuse of data across various applications, so that tomorrow's programs can share and process data even when these programs have been designed totally independently.

Structured data

Digital documents allow, in principle, much richer automated processing - content selection, information extraction, price comparisons, or document clustering. To facilitate this, data is structured; documents are given internal structure. There are many different formats for structured data, some simple in their description, others complex and rich. Many communities (biologists, engineers, geologists, businesses) are designing new formats to allow them to put machine-understandable data on the Web. This will allow data to be shared and processed by automated tools as well as by people.

Relational Data

Relational databases represent information by representing relations between entities (such as the relation between book and author, or the relation between book and publisher). Rather than exchange whole databases, query languages, such as SQL, allow users to retrieve, from the database, information about specified entities. Relational databases are appropriate for representing uniformly structured data, where all entities of a given type can be represented by specifying a given collection of relations (every part has a price, every employee has a manager). But they are ill adapted to the open-ended nature of information on the web, where we may find that one manager also plays in a band, and so have to represent the fact that she plays the saxophone.

Markup

One common form of structuring is markup. Markup originated as a means of structuring text; before the computerization of the printing industry, markup was annotation written by a copy editor on a manuscript, to indicate structure (chapter, heading, paragraph...) and style (italic, bold, etc.). Markup now refers to sequences of characters, known as tags, inserted in a text or word processing file. The original use of markup was to indicate how the file should look when printed or displayed. Markup is now often used to describe a document's logical

structure, or as a format for describing an abstract logical structure, so-called 'semi-structured data' - a replacement for the representation of structured data by traditional relational databases.

Standard frameworks for markup, such as SGML (Standard Generalized Markup Language) and XML (eXtensible Markup Language), have been adopted by many metadata initiatives, including the semantic web, where XML markup is used to structure the metadata attached to a document. An example of SGML markup is the HyperText Markup Language (HTML), *lingua franca* of the World Wide Web.

A general markup language, such as XML, can be applied to encode content and structure for applications that go far beyond the original purposes of markup for typesetting and display. Markup can be used to tag entities in a document (addresses, prices, or names), to tag logical roles (author, publisher), to tag logical connections, for example linking a price to an entity, or to tag phrases and parts of speech, in order to indicate a text's detailed grammatical structure. Applications of XML are found everywhere: in bioinformatics and linguistics, in business-to-business applications, in cataloguing and indexing, and in scholarly annotation of ancient texts.

Markup languages define the scope of what can and cannot be expressed in markup. For example, XML provides controlled flexibility, and allows us to represent semi-structured data that does not fit well in the relational mould. XML tags come in pairs, <author>...</author>, which act as "named brackets". These brackets must be properly nested, so an XML document has a hierarchical structure in which each layer of the hierarchy consists of text, interspersed with elements from the layer below tagged with names (such as "author"). This allows a mixture of structure and free text.

Most applications require further restrictions. For example, HTML is defined by a document type definition (DTD) that specifies which elements may occur in an HTML document (headings, paragraphs, lists...), and structural rules (for example, list elements occur within lists) that an HTML document must follow. A DTD can specify as many constraints on the structure as are needed for a particular application, or as few. So an author element might require a surname, allow any number of forenames, and permit nothing else.

From a scientific perspective, XML and the structures it allows us to express represent just one of many possibilities for structuring data. The science underlying XML provides an understanding of the ways in which data may be structured. It provides query languages and algorithms for retrieving data in response to a query. It provides the conceptual framework for understanding how structure may be specified, for example by a DTD, and algorithms for checking that a document conforms to a specified form. It provides the basis for assurances of the integrity and provenance of data, and so on.

Processing and Communication

Processing is the transformation of data from one form to another. Information is data interpreted, organized and structured. For example, the English documents on the web form a large dataset; when we use a search engine to count the numbers of documents containing a given search term it finds on the web, we extract information from this data. The ability to collect, aggregate, and organize data allows us to create and represent information. Knowledge is information that has been analysed so that inter-relationships are identified, formalized, and represented. Thus processing can extract information from data, and transform information into knowledge. These results must then be communicated effectively to the user.

Natural Language Processing

Today (2002), machines are widely used for document retrieval. Future software agents will use Natural Language Processing (NLP) tools to extract relevant information from documents in response to user queries, create summaries tailored to the user's needs, and collate, assemble and present information derived from a multitude of sources.

Human readers see structure in texts: words, sentences, paragraphs, documents.... They attribute meanings to documents, and structure these meanings - as facts, arguments and conclusions, and so on. Representing, processing and communicating structures and meanings, in ways that make these easily accessible to machine processing, so that users can easily access information relevant to their needs, are key issues for informatics.

Automated text-processing tools can tag parts of speech, and annotate grammatical links: the connection between verb, subject, and object; the connection between a pronoun and the phrase it refers to. Deeper linguistic processing can disambiguate word senses. Such tools are components of natural language understanding. They convert texts into machine-accessible sources of information, and provide the basis for a variety of information processing applications.

For example, current document retrieval systems, using keyword search, allow users to find documents that are relevant to their needs, but most leave it to the user to extract the useful information from those documents. Users, however, are often looking for answers, not documents. Information extraction tools, based on natural language understanding, find and extract information from texts. Such tools, already used, by intelligence analysts and

others, to sift through large amounts of textual information, will become commonplace. Document clustering is another application that draws on natural language technology. By examining patterns of word occurrences, together with syntactic and semantic structures, it is possible to cluster documents by topic, or to search for documents similar to a given example.

Automated natural language generation is also being applied, to new forms of information delivery. Machine-generated text can be used to present information tailored to the user (Oberlander et al. 1998). Such tools will be applied to present information derived from database queries, information extraction, and data mining. Tools for document clustering will be linked to natural language generation to provide automatically generated summaries drawing on a variety of sources.

Natural language processing is one example demonstrating the way informatics relates to longer-established disciplines. It relies on computer science for underlying software and hardware technologies, and for algorithms that make this processing feasible. It draws on logic and linguistics for appropriate representations of linguistic and semantic structures, on machine learning techniques from artificial intelligence for tools that extract from large text corpora information on the words and concepts relevant for a particular domain, and on cognitive science and psychology for an understanding of how people process and react to information.

These disciplines are drawn together, in informatics, by the common purpose of understanding how language can communicate information between human users and a formal representation stored in a machine. This marriage of computational and theoretical linguistics with cognitive psychology and neuroscience has generated new tools, and also thrown new light on human communication.

It is clear that information, and hence informatics, must play a pivotal role in any analysis of human communication. Informatics is also transforming other areas of science: it provides a new paradigm for analysing complex systems, as compositions of simpler subsystems that process and communicate information. Long-standing challenges to scientific analysis, are being transformed by the new paradigm.

For example, in biology, informatics provides not just tools for data-processing and knowledge discovery, but also a conceptual framework for studying the information stored and communicated by genes, and processed by biochemical cycles that ‘run the genetic program’ to produce structure and form.

In cognitive science it provides the conceptual tools needed to develop models of the connection between cognition and the observed structure and function of the billions of neurons that make up the brain. It also provides the technologies that allow us to observe and analyse the structure and operation of the living brain in ever greater detail, and to test our models by simulation of very much simplified subsystems, which, despite their relative simplicity, are complex beyond analytical analysis.

Looking forward

The technologies underlying the digital storage, processing and communication of information are improving relentlessly. Since 1965 these improvements have followed ‘Moore’s Law’: for a given price, both processor speeds and memory capacity double every 18–24 months (Moore 1965). Communication bandwidth follows a similar pattern of growth, but doubles in 12 months or less (Poggio 2000). These rates of exponential growth are predicted to continue, so processing speed and memory capacity will increase by a factor of 100, and communication bandwidth by a factor of 1000 or more, every ten years.

The combination of digitisation and global connectivity makes data available in unprecedented volume. It is estimated that humanity creates more than an exabyte of data each year (Lyman and Varian, 2000). Nevertheless, it will soon be technologically possible for an average person to access virtually all recorded information. The availability of cheap processing will make it increasingly feasible to restructure data into knowledge on demand. New technologies are being developed to automatically organize this material into forms that can help people quickly and accurately satisfy their information needs, realizing, and surpassing, Vannevar Bush’s prescient vision of the ‘Memex’ – ‘a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility’ (Bush 1945). These technologies are indeed ‘creating a new relationship between thinking man and the sum of our knowledge’.

By 2025, if Moore’s law continues to apply, we will have, in our pockets and on our desktops, computers that each have the raw computing power of the human brain, computers linked to each other by a telepathic communication network. We currently have little idea of how we might structure and program such devices to achieve what we humans find straightforward, but already today’s machines extend our capabilities by performing tasks we find impossible. We can be sure that technological changes will continue to revolutionize the ways we manage, share, and analyse data, and will provide new ways of transforming data into information and knowledge.

References

- Bauer, W.F. (1996) 'Informatics and (et) Informatique.' *Annals of the History of Computing*, 18(2).
<http://www.softwarehistory.org/history/Bauer1.html>
- Berners-Lee, Tim. (1998) Semantic Web Roadmap. <http://www.w3.org/DesignIssues/Semantic.html>
- Bush, V. (1945) 'As we may think', *Atlantic Monthly* 176(1) (July), pp. 101—8.
<http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm>
- Dreyfus, Ph. (1962) 'L'informatique.' *Gestion*, Paris, Juin 1962, pp. 240—1.
- Frieden, B.R. (1998) *Physics from Fisher Information: a Unification*, Cambridge University Press
- Google (2002) <http://www.google.com/search?q=informatics>
- Gorn, S. (1983). 'Informatics (computer and information science): Its ideology, methodology, and sociology', in F. Machlup & U. Mansfield (eds.), *The study of information: Interdisciplinary messages*, pp. 121—40.
- Greenes, R.A. and Shortliffe, E.H. (1990) Medical Informatics: An emerging discipline with academic and institutional perspectives. *Journal of the American Medical Association*, 263(8) pp.1114—20.
- Lesk, Michael. (1997) How Much Information Is There In the World? <http://www.lesk.com/mlesk/ksg97/ksg.html>
- Lyman, P and Varian, H. (2000) How Much Information? *The Journal of Electronic Publishing*, 6(2). ISSN 1080-2711
<http://www.press.umich.edu/jep/06-02/lyman.html>
- Machlup, F. and Mansfield U. (eds.), (1983) *The study of information: Interdisciplinary messages*. John Wiley & Sons.
- Mikhailov, A.I., Chernyl, A.I., and Gilyarevskii, R.S. (1966) Informatika – novoe nazvanie teorii nauènoj informacii. *Nauèno tehnièskaja informacija*, 12, pp 35—9.
- Mikhailov, A.I., Chernyl, A.I., and Gilyarevskii, R.S. (1967) Informatics – new name for the theory of scientific information. *FID News Bull.* 17(2), pp. 70—4.
- Moore, G.E. (1965) Cramming more components onto integrated circuits. *Electronics*, 38(8) pp. 114—7.
<http://www.intel.com/research/silicon/moorespaper.pdf>
- Oberlander, J. O'Donnell, M. Knott A. and Mellish, C. (1998). Conversation in the museum: experiments in dynamic hypermedia with the intelligent labeling explorer. *New Review of Hypermedia and Multimedia*, 4, pp. 11—32.
- Oxford English Dictionary (1989) second edition. Oxford University Press.
- Poggio, A. (2000) Information and Products, in *Eaglehart's Colloquium, the unfinished revolution*. Stanford
http://www.bootstrap.org/colloquium/session_10/session_10_poggio.jsp

Further Reading

- Association for Computational Linguistics (ACL) <http://www.aclweb.org/>
- Cole, Ron *et al.* (eds), (1998) *Survey of the State of the Art in Human Language Technology*. Studies in Natural Language Processing, Cambridge University Press; ISBN: 0521592771. <http://cslu.cse.ogi.edu/HLTSurvey>
- Cover, Robin (Ed). (2002) *XML Cover Pages*, <http://www.oasis-open.org/cover/sgml-xml.html>
- Graves, J.R. and Corcoran, S. (1989) 'The Study of Nursing Informatics', *Image: Journal of Nursing Scholarship*, 21, pp. 227—31 <http://www.nih.gov/ninr/research/vol4/Overview.html>
- Martin, W.J. (1988) *The Information Society*, London: ASLIB.
- Musen, Mark A. (1999) 'Stanford Medical Informatics: Uncommon research, common goals'. *MD Computing*, January/February 1999, pp.47—50. <http://camis.stanford.edu/MDComputing.pdf>

Shortliffe, E.H. Perreault, L.E. Wiederhold, G. and Fagan, L.M. (1990) *Medical Informatics: Computer Applications in Health Care*, Addison-Wesley.

Social Informatics <http://www.slis.indiana.edu/si/concepts.html>

Text Retrieval Conference (TREC) <http://trec.nist.gov/>

See also: **artificial intelligence; communication; communication and IT; computer science; data modelling; database; database management; human-computer interaction; hypertext; indexing; information retrieval; information science; intelligent agents; knowledge management; knowledge-based systems; machine translation; mark-up languages; metadata; neural network; relational database; information systems; search engines; software; string indexing; World Wide Web.**

MICHAEL P. FOURMAN

(4107 words)